LECTURE #4: (AGNOSTIC) PAC LEARNING

Instructor: Aditya Bhaskara *Scribe: Ishaan Rajan*

CS 5966/6966: Theory of Machine Learning

January 20th, 2022

Abstract

In this lecture we introduce the notion of agnostic PAC learning and the idea that finite classes are PAC learnable.

1 REVIEW AND INTRODUCTION

Last class, we proved the No Free Lunch theorem, which had two main takeaways:

- There is no universal learning algorithm, even if allowed to be inefficient
- If we wish to learn some arbitrary function over a set of m points, at least $\frac{m}{2}$ training examples are needed

In other words, in the PAC model, the hypothesis class $\mathcal{H}$ – which consists of all the possible functions over the domain – cannot be learned.

Recall Valient's definition of learnability:

- A concept class is *l*earnable if there exists an efficient algorithm $\mathcal{A}$ with the following property: for all $\epsilon > 0$, there exists m number of samples such that when given m i.i.d. samples from $\mathcal{D}$ along with their labels, $\mathcal{A}$ produces a hypothesis h with risk less than ϵ, with probability ≥ 0.9, where risk is the expected error on sample from distribution.

2 PAC LEARNING – REALIZABLE CASE

A concept class $\mathcal{H}$ is PAC learnable over domain X (the realizable case) if there exists an algorithm $\mathcal{A}$ that for all $\epsilon, \delta > 0$ and distribution $\mathcal{D}$, has the following properties:

$\mathcal{H}$: a bunch of hypothesis over X
Note: A is allowed to be inefficient.

- Given $m(\epsilon, \delta)$ – the sample size which does not depend on $\mathcal{D}$ – samples $(x, (f(x))$, where $x \sim \mathcal{D}$ and f is an unknown function in $\mathcal{H}$, it outputs h with risk at most ϵ with probability at least $1 - \delta$.

As such, we can conclude that h need not belong to $\mathcal{H}$. This is known as improper learning.

Essentially our goal here is to find a true label function $f \in \mathcal{H}$, and declare success if we find h that has risk $\leq \epsilon$.

3 PAC Learning – non-realizable case

A concept class $\mathcal{H}$ is agnostically PAC learnable over domain X (the realizable case) if there exists an algorithm $\mathcal{A}$ that for all $\epsilon, \delta > 0$ and distribution $\mathcal{D}$, has the following properties:

- Given $m(\epsilon, \delta)$ samples $(x, (f(x))$, where $x \sim \mathcal{D}$ and f is an unknown function not necessarily in $\mathcal{H}$, it outputs h with risk at most ϵ more than the risk of the h' in $\mathcal{H}$ that is "closest" to f with probability at least $1 - \delta$.

This weakens inductive bias

- Again, the sample size need not depend on $\mathcal{D}$, and h need not belong to $\mathcal{H}$.

4 Every finite class is PAC learnable (even agnostic)

Suppose $\mathcal{H}$ has only finitely many hypothesis $h_1, h_2, \cdots h_N$ over (possibly infinite) input space X. We can prove that $\mathcal{H}$ is still PAC-learnable with the following generic algorithm: Empirical Risk Minimization (ERM).

At least $\frac{1}{\epsilon}$ samples are needed

- get m examples
- find $h \in \mathcal{H}$ that minimizes empirical risk
- output h (we are guaranteed one such h)

Empirical risk of h and sample S is defined as

$$\frac{1}{|S|} \sum_{x \in S} 1_{h(x) \neq f(x)}$$

Essentially minimizing training error here

1 *Question*. What is the difference between an example and a sample?
An example is one such $(x, f(x))$, and a sample $\mathcal{S}$ is a collection of these examples.
2 *Question*. When is ERM bad?
If there are too few examples (i.e. m is too small), or we just got unlucky with examples, ERM will not perform well.

5 Representative Sample

Here we will introduce the definition of a representative sample which will be further explained in the following lecture.
Let $\mathcal{H}$ be a hypothesis class and X be an input space with a distribution $\mathcal{D}$ on it, and let f be a target function. Sample $S \subseteq X$ is said to be ϵ - "representative" if for all h in $\mathcal{H}$, we have:

$$|\frac{1}{|S|} error(S, h) - risk_D(h, f)| < \epsilon$$

In other words, we have the empircal risk minus the true risk with respect to $\mathcal{D}$.